ARTICLE

CAG Expansion in the Huntington Disease Gene Is Associated with a Specific and Targetable Predisposing Haplogroup

Simon C. Warby,[1] Alexandre Montpetit,[2] Anna R. Hayden,[1,2] Jeffrey B. Carroll,[1] Stefanie L. Butland,[1] Henk Visscher,[1] Jennifer A. Collins,[1] Alicia Semaka,[1] Thomas J. Hudson,[2,3] and Michael R. Hayden[1,*]

Huntington disease (HD) is an autosomal-dominant disorder that results from ≥36 CAG repeats in the HD gene (*HTT*). Approximately 10% of patients inherit a chromosome that underwent CAG expansion from an unaffected parent with <36 CAG repeats. This study is a comprehensive analysis of genetic diversity in *HTT* and reveals that HD patients of European origin (n = 65) have a significant enrichment (95%) of a specific set of 22 tagging single nucleotide polymorphisms (SNPs) that constitute a single haplogroup. The disease association of many SNPs is much stronger than any previously reported polymorphism and was confirmed in a replication cohort (n = 203). Importantly, the same haplogroup is also significantly enriched (83%) in individuals with 27–35 CAG repeats (intermediate alleles, n = 66), who are unaffected by the disease, but have increased CAG tract sizes relative to the general population (n = 116). These data support a stepwise model for CAG expansion into the affected range (≥36 CAG) and identifies specific haplogroup variants in the general population associated with this instability. The specific variants at risk for CAG expansion are not present in the general population in China, Japan, and Nigeria where the prevalence of HD is much lower. The current data argue that *cis*-elements have a major predisposing influence on CAG instability in *HTT*. The strong association between specific SNP alleles and CAG expansion also provides an opportunity of personalized therapeutics in HD where the clinical development of only a small number of allele-specific targets may be sufficient to treat up to 88% of the HD patient population.

Introduction

Huntington disease (HD [MIM 143100]) is a dominantly inherited progressive neurodegenerative disorder that results from a mutation that expands the polymorphic trinucleotide (CAG) tract in *HTT*. The average control CAG tract size in the general population is 17–20 repeats.[1] However, in HD patients, one of the two copies of the gene has a CAG tract that has expanded to 36 repeats or more.[2]

The size of the CAG tract can be unstable and has a bias toward increasing size, especially when transmitted by the male germline.[3] Initially, the new mutation rate for HD was believed to be very low, and the disease was limited to families with a history of HD. Recent estimates have found that CAG expansion into the disease range is more common than initially anticipated and the new mutation rate for HD could be 10% or greater.[4,5]

Many factors are believed to contribute to CAG instability, including the size of the CAG tract, CAG tract interruptions, sex and age of the transmitting parent, environmental influences,[6,7] and other genetic *cis*-elements and *trans*-factors.[8,9] Although a larger CAG tract size and transmission through the male germline have been clearly demonstrated to increase CAG instability, *trans*-factors such as DNA repair machinery are also believed to play an important contributing role.[10] For instance, the CAG instability of HD transgenic mice was rescued when crossed with mice lacking either the mismatch repair enzyme *MSH2* (MIM 609309)[11] or the base excision repair enzyme *OGG1* (MIM 601982).[12] Although *cis*-elements are known to modify CAG instability in other genes,[13] previous data have argued against a role for *cis*-elements in CAG instability in *HTT*.[14,15]

Numerous studies have investigated the origins of the disease by constructing haplotypes of the *HTT* region in specific ethnic populations.[15–30] These studies have been limited to small sets of allelic markers because, aside from the CAG tract, only a few polymorphisms in *HTT* have been previously characterized.[31] Many of these studies found positive correlations between specific markers and disease chromosomes[27] and concluded that the HD mutation originates from a common descent, although not necessarily from a single founder.[26–28]

To our knowledge, this is the first comprehensive analysis of genetic diversity in *HTT* on HD (>36 CAG), intermediate allele (27–35 CAG), and control (<27 CAG) chromosomes. Despite the genetic diversity around *HTT* in the general population, we identify a subset of tagging single nucleotide polymorphisms (SNPs) that are highly associated with disease chromosomes (≥36 CAG) of European descent. We report many SNPs that are highly sensitive markers of disease chromosomes and have stronger linkage associations with CAG expansion than any previously characterized polymorphism in the gene region. The disease-associated SNPs constitute a cluster of similar haplotypes (haplogroup A) found on 95% of disease chromosomes.

[1]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, Vancouver, BC V5Z 4H4, Canada; [2]McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 1A4, Canada; [3]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada
*Correspondence: mrh@cmmt.ubc.ca
DOI 10.1016/j.ajhg.2009.02.003.

Furthermore, we find the same haplogroup A is significantly enriched on unaffected chromosomes with an elevated CAG size (27–35 CAG; intermediate alleles) relative to the general population (15–26 CAG). The presence of CAG-expanded chromosomes on a single haplogroup supports the hypothesis that predisposing *cis*-elements are a major determinant of CAG instability.

The strong association between specific SNPs and CAG-expanded chromosomes now provides an attractive therapeutic opportunity for the treatment of HD. The association of specific SNPs combined with high rates of heterozygosity in HD patients provides suitable targets for allele-specific knockdown of the mutant gene product.

Material and Methods

Sample Populations

A total of 190 SNPs were identified and verified in the *HTT* region on chromosome 4p16.3 through a combination of data from direct sequencing and phase I/II HapMap Data Coordination Centre (Figure S1 available online). Direct sequencing was performed on Canadians of European origin, from the clinical database at the University of British Columbia (UBC) DNA and Tissue Bank for Huntington Disease Research (procedures followed and patient consent were in accordance with UBC ethics H05-70532-A003). The HapMap cohort includes 30 trios (mother, father, offspring) from U.S. residents with northern and western European ancestry (originally collected by the Centre d'Etude du Polymorphisme Humain), 30 trios from the Yoruba people of Ibadan, Nigeria, and 45 unrelated individuals from both Beijing, China and Tokyo, Japan.

DNA from 65 HD patients was included in this study (38 male/27 female). 30 samples were selected because of trio information availability and 35 were selected randomly from the UBC database. A second cohort of 203 HD patients was used as a replication cohort to confirm the SNP associations. The control (unaffected spouse) group consisted of 58 randomly selected subjects (19 male/29 female). The intermediate allele population have 27–35 CAGs in *HTT* and consisted of 66 subjects (30 male/36 female). The cohorts were comprised of mostly Canadians of European descent, with a small proportion having indigenous Canadian ethnicity (<5%).

Selection and Genotyping of SNPs in HapMap Cohorts

Selection of SNPs was achieved through use of direct sequencing results, online SNP databases such as the SNP Consortium and dbSNP, as well as data from Phase I and II of the International HapMap Project. A total of 190 SNPs were identified within the *HTT* region (Figure S1) and subsequently genotyped in the HapMap population.[32,33]

Phylogenetic and Linkage Analysis

Phylogenetic analysis of genotypes was performed with the Mega3 software.[34] Each individual from the HapMap cohort was compared based on sequence similarity at the 190 SNP positions to construct a neighbor-joining tree rooted on the chimpanzee sequence. Linkage disequilibrium (LD) analysis was performed with Haploview software.[35] Linkage plots are displayed with the standard LD color scheme (D'/LOD) where color represents linkage, in decreasing order: red (LOD $\geq$ 2 and D' = 1), pink (LOD $\geq$ 2 and D' < 1), blue (LOD < 2 and D' = 1), and white (LOD < 2 and D' < 1).

Determination of tSNP Genotypes and Phasing in HD Patients and Controls

A panel of 22 nonredundant tSNPs were selected as a means to estimate haplotypes in the HD patients, 27–35 CAG individuals, and unaffected spouse controls from the general population. Tag SNPs were selected with the Tagger program.[36] Genotyping and phasing of HD, 27–35 CAG, and controls was performed with the SNPstream genotyping assay.[37] The methods in brief are described. (1) Primers for PCR and SNP-IT were designed based on SNP loci and its appropriate sequence with web-based Autoprimer software. (2) Oligonucleotide microarrays were produced in a 384-well plate set-up on a glass-bottomed plate. (3) Genomic sequences carrying the SNPs of interest were amplified in a 12-plex PCR reaction. (4) PCR clean-up was performed. (5) Post-PCR cycle extension for the SNP-IT reaction. (6) Assay results were interpreted by two-color fluorescence on SNPstream UHT Array Imager. (7) Data analysis, quality control, and graphical interpretation of results were performed on SNPstream software. Phasing was performed by incorporating genotype information of family trios into PHASE 2.0 software. Direct sequencing of the CAG tract was performed with primers HTTg010_F (ATTACAGTCT CACCACGCCC) and HTTg010_R (GACAAGGGAAGACCCAAGTG).

Haplogroups

Individual haplotypes were combined into haplogroups with specific tSNPs to define the groupings. Haplogroup A was defined with tSNPs that had allele distributions that were significantly associated with disease chromosomes (chi-square disease versus control) and had high sensitivity (>0.95). Haplogroups B and C were defined manually to classify the remaining haplotypes. The remaining unclassified haplotypes were singleton or rare haplotypes and were put into an "other" haplogroup but are not necessarily related to each other. Analysis of the haplogroup variants was performed only on chromosomes that were defined as haplogroup A. Variants A1–A5 were defined and distinguished from each other with the remaining tSNPs not used in the definition of haplogroup A.

Statistics

The association of SNPs with CAG expansion was determined via chi-square test on the allele frequencies with a bonferroni correction when appropriate. Odds ratio is the fold increase of having the disease if the specific allele is present ((Ad/An case)/(Ad/An control)), where Ad = disease-associated allele; An = nondisease. Sensitivity is a measure of how well a SNP allele correctly identifies the disease state (proportion of cases with the disease-associated allele (Ad/Ad+An case)). Specificity is a measure of how well a SNP allele correctly identifies the control state (An/Ad+An control). Positive predicted value (PPV) is the proportion of subjects with the disease-associated allele who have the disease (Ad case/Ad case + Ad control). Negative predicted value (NPV) is the proportion of subjects who do not have the disease-associated allele and do not have the disease (An control/An case + An control). When indicated, a Student's t test was performed for comparison between two groups. One-way ANOVA was performed for more than two groups with Tukey's post-hoc comparison.

Results

SNP Validation and Characterization

A total of 190 SNPs were identified and verified in the *HTT* region on chromosome 4p16.3 (Figure 1A). Twelve SNPs are

Figure 1. Genetic Diversity Is Observed in the Human *HTT* Region

(A) Schematic representation of the location of 190 SNPs (red triangles) relative to the intron/exon structure of *HTT*. SNP 1 is located 5′ of the gene, whereas SNP 190 is the furthest 3′ of the gene.

(B) Each of the 190 SNPs has a minor allele frequency (MAF) and the number of times each MAF occurs is plotted to demonstrate that there is a large amount of diversity in the *HTT* region in the HapMap populations. 23% of SNPs were very common (MAF $>$ 0.20). Many (45%) rare SNPs (MAF $<$ 0.05) were also detected.

(C) The MAF of each SNP is aligned for each ethnic population in the HapMap cohort. The pattern is similar for each ethnic group, although the Yoruba have the most SNPs overall and tend to have higher MAFs.

(D) Plot of linkage disequilibrium (LD) in the *HTT* region in the HapMap cohorts. Linkage across *HTT* is relatively high (red is high LD, white is low LD), enabling the use of tagging SNPs to sample the HD population. Chinese and Japanese have similar linkage and are pooled into an Asian grouping. As expected, in both the MAF and the LD plot, the greatest diversity is seen in Yoruba relative to Europeans and Asians.

in the promoter region, 11 exonic (coding), 157 intronic, and 9 in the 3′ UTR region of *HTT*. For each SNP, the minor allele frequency (MAF) was determined from the pooled HapMap populations. A large proportion (23%) of the 190 SNPs in the *HTT* region were common, having a MAF of >0.20 (Figure 1B). These common SNPs will therefore comprise the majority of diversity in the gene and heterozygous individuals in these populations. The abundance of

common SNPs is important because it may present the opportunity for allele-specific targeting of the disease-carrying transcript. We also detected many rare SNPs (45% of all SNPs) with a MAF < 0.05, a finding that is consistent with previous studies (46% of SNPS with MAF < 0.05 in ENCODE regions[32]).

To compare the genetic diversity of the *HTT* region between ethnic groups, the MAF (Figure 1C) and pattern of LD (Figure 1D) was plotted for each HapMap population. Although there are clear similarities in the pattern of MAFs between the populations, the Yoruban population exhibited the most overall diversity, as expected.[32,38,39] In Yoruba, the MAFs tend to be higher (and are therefore more frequently heterozygous), and rare SNPs (MAF < 0.05) were observed more often (Figure 1C). The amount of linkage disequilibrium (LD) across the gene region was less in the Yoruban population relative to the other groups (Figure 1D). Phylogenetic analysis of individuals in the HapMap cohort is consistent with a substantial amount of diversity in the *HTT* region, with the greatest diversity in Yoruba (Figure S2). In contrast, the Asian group has fewer SNPs overall, lower MAFs (Figure 1C), and highest LD (Figure 1D) relative to the other groups.

SNP Frequency Distribution on HD Chromosomes

With the patterns of LD in the 190 validated SNPs, a panel of 22 nonredundant tagging SNPS (tSNPs)[40] was designed to efficiently assess genotypes and construct haplotypes in the *HTT* region. In order to determine which genotypes came from each chromosome within each individual, chromosomes were phased by CAG tract size. This allows grouping of chromosomes by CAG tract size and comparison of genotypes between the groups. HD patients ($n = 65$) of European ancestry were genotyped at the 22 tSNP positions. Genotype information for disease chromosomes (>35 CAG) and control chromosomes (<36 CAG) from these HD patients is shown in Figure 2A. Allele counts for the major (A) and minor (B) alleles are given at each of the 22 tSNP positions.

HD chromosomes have a dramatically different allele distribution compared to control chromosomes. Of the 22 tSNPs, more than half (12 SNPs) are significantly associated with disease chromosomes (chi-square $p < 0.0023$; alpha $= 0.0023$ after bonferroni correction). For example, at tSNP # 11, the allele distribution on the control chromosome (A-32, B-33) is significantly different compared to the disease chromosome (A-1, B-64) (chi-square $p = 1.7 \times 10^{-10}$). This significant association is matched with an extremely high odds ratio (OR = 66) at this and other SNP positions.

In addition to many tSNPs being significantly associated with disease chromosomes, some tSNPs have a single allele that is a highly sensitive marker of disease chromosomes (Figure 2). Sensitivity is a measure of how frequently a specific allele occurs on a disease chromosome, and for 19 tSNPs the sensitivity is >0.85. It is important to note that not all significantly associated SNPs are highly sensitive markers of disease chromosomes. For example, tSNP #11 is a highly sensitive marker of disease chromosomes ($64/65 = 0.98$) whereas tSNP #182 is not ($34/65 = 0.52$). tSNP #182 is, however, a specific marker (specificity 0.94) because of its low MAF on general population chromosomes.

These data were confirmed in a replication cohort of 203 HD patients. The same tSNPs were significantly associated with the disease chromosome (Figure S3). Surprisingly, some tSNPs are not associated with disease chromosomes, even in the combined cohort of 268 HD patients. For example, at tSNP #63 the allele distributions between disease (A-29, B-239) and control (A-14, B-254) chromosomes remain similar despite the fact that surrounding tSNPs have extremely strong associations. The incomplete allelic association and variable marker sensitivity is therefore not consistent with a simple single-founder hypothesis for the origin of HD chromosomes. This unusual pattern across the gene region would require an unlikely series of recombination events or significantly different mutation rates at different SNP positions.

SNP Frequency Distribution on 27–35 CAG Chromosomes

Individuals with increased CAG tract size (27–35 CAG) relative to the general population are not at risk for developing the signs and symptoms of HD. However, children of parents with 27–35 CAG may be at risk for inheriting a CAG allele that has expanded in the disease range.[8] New mutations for HD have been shown to originate from this 27–35 CAG population. Previous to this study, it has not been clear whether the haplotype of 27–35 CAG chromosomes is most similar to control chromosomes in the general population or whether they have genetic similarities to chromosomes that have already expanded into the disease range.

In total, 66 chromosomes in the 27–35 CAG range were phased and compared to 116 control chromosomes from the general population (<27 CAG) at each of the 22 tSNP positions (Figure 2B). The frequency of alleles on 27–35 CAG chromosomes was significantly different from control chromosomes at 11 tSNP positions (chi-square $p < 0.0023$). The odds of these specific tSNPs being associated with CAG expansion is high (odds ratio ≥ 3). Notably, these 11 out of 12 tSNP positions are also significantly associated with HD chromosomes.

Many tSNPs are also sensitive markers of 27–35 CAG chromosomes. Eighteen tSNPs have a sensitivity ratio >0.85. Notably, these 18 out of 19 tSNP positions were also sensitive markers of disease chromosomes. Furthermore, similar to HD chromosomes, not all tSNPs with significant associations have a single allele that is a sensitive marker of 27–35 CAG. Again, tSNP #182 is significantly associated with 27–35 CAG chromosomes (chi-square $p = 1.7 \times 10^{-07}$) but not a sensitive marker ($30/66 = 0.45$).

Control chromosomes all had similar allele frequency distributions, regardless of the source of the control chromosome. There were also no significant differences in the allele frequencies (chi-square $p > 0.0023$) on control chromosomes

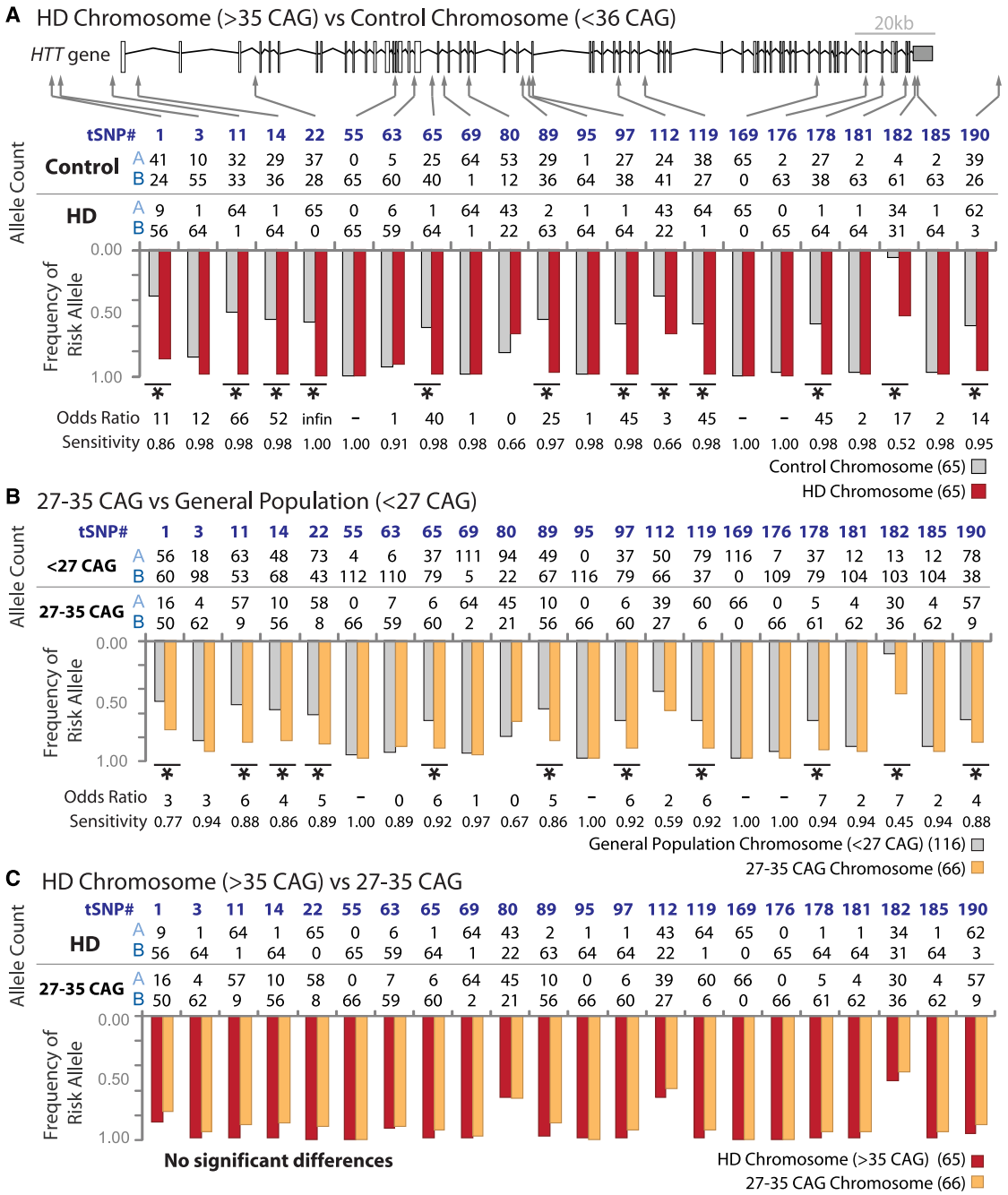

Figure 2. Specific SNPs Are Highly Associated with CAG-Expanded Chromosomes

(A) HD patient chromosomes were phased to allow comparison between the disease chromosome (>35 CAG) and control chromosome within each patient (total 65 individuals). tSNP is identified by number and its position indicated relative to *HTT*. Alleles are either A or B (major or minor). Allele counts are indicated (middle) and the frequency graphed (below). Twelve out of 22 tSNPs have a significantly different allele distribution between HD and control chromosomes (*chi-square $p < 0.0023$).

(B) Allelic frequency on 27–35 CAG chromosomes is similar to disease chromosomes. Allele counts are indicated for phased control chromosomes ($n = 116$) and compared to 27–35 CAG chromosomes ($n = 66$) that contain an intermediate CAG tract size for *HTT* and may result in new mutations for HD in future generations. Eleven out of 22 tSNPs have a significantly different allele distribution between 27–35 CAG and control chromosomes (*chi-square $p < 0.0023$). These 11 associated tSNPs were found in both HD and 27–35 CAG chromosomes and appear to be common on CAG-expanded chromosomes.

(C) There is no significant difference in the allele distribution between 27–35 CAG and HD chromosomes for any tSNPs.

(<27 CAG), whether they came from the general population or the control chromosome (lower CAG) from HD patients or control chromosomes of 27–35 CAG carriers (Figure S4).

The allele frequencies on 27–35 CAG chromosomes were not significantly different at any tSNP position (chi-square $p > 0.0023$) from HD chromosomes (Figure 2C). Taken

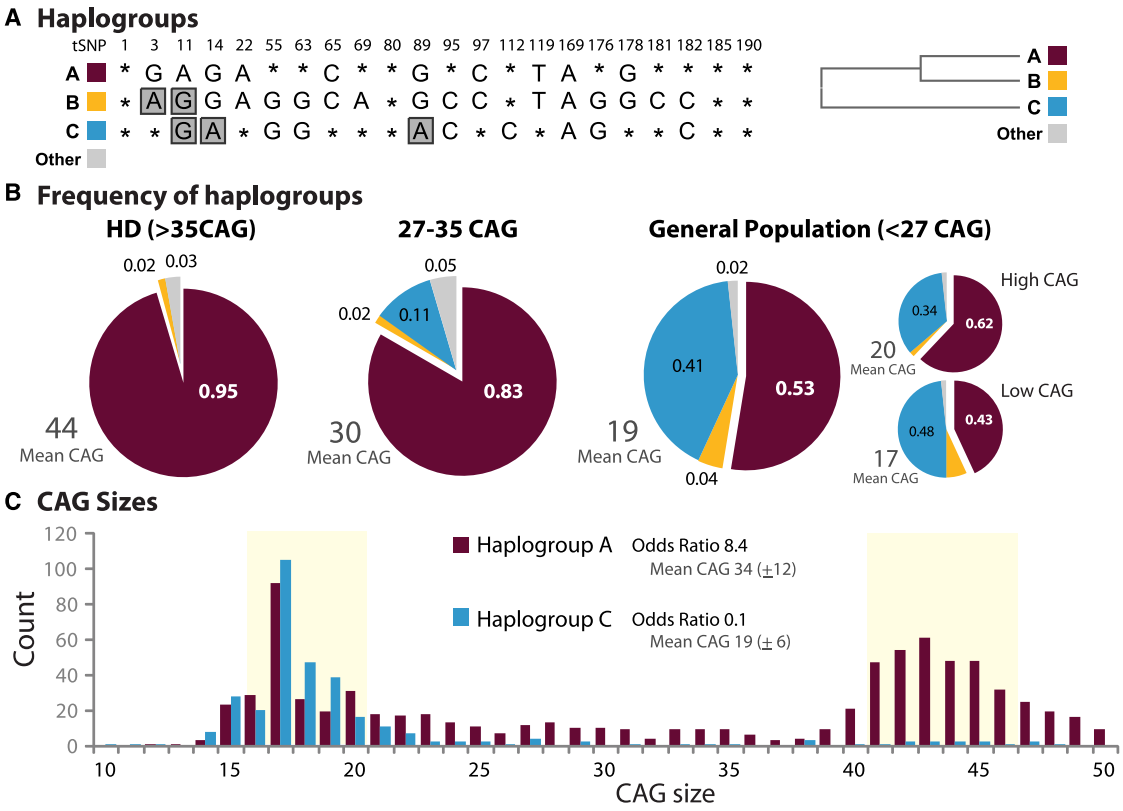

Figure 3. CAG-Expanded Chromosomes Are Associated with Haplogroup A

(A) Three major haplogroups (A, B, C) are defined with 22 tSNP positions. The nucleotide defining each haplogroup at each tSNP is shown. Variable tSNP positions are indicated (*). tSNPs with a gray box indicate nucleotide changes relative to haplogroup A. The amount of similarity between the haplogroups is indicated by a neighbor-joining tree (right).

(B) Frequency distribution of haplogroups on HD (n = 65), 27–35 CAG (n = 66), and general population (n = 116) chromosomes. CAG-expanded chromosomes (>27 CAG) are enriched for haplogroup A relative to the general population. Chromosomes from the general population with <27 CAG phased for CAG size (right) demonstrate that high-normal CAG chromosomes also have an enrichment for haplogroup A relative to low-normal CAG chromosomes. The mean CAG tract size for each group is indicated.

(C) CAG size distribution for all chromosomes on haplogroups A or C. In the chromosomes used in this study, the mean CAG sizes for haplogroup A are significantly larger ($p < 0.00001$, t test) than haplogroup C. The high odds ratio on haplogroup A is an indication that CAG expansion is much more likely to occur on haplogroup A chromosomes.

together, these data suggest that as a group, 27–35 CAG chromosomes and HD chromosomes are genotypically similar across the *HTT* region.

Haplogroup Frequencies on CAG-Expanded Chromosomes

Haplogroups were defined manually with tSNPs that are significantly associated ($p < 0.0023$) and are highly sensitive markers (>0.95) of disease chromosomes (Figure 3A). Three major haplogroups (A, B, and C) could be used to describe >96% of all chromosomes in our study cohort of HD patients, 27–35 CAG individuals, and controls from the general population. A neighbor-joining phylogeny demonstrates that haplogroups A and B are much more closely related than either are to haplogroup C (Figure 3A). The "other" haplogroup comprised singletons that could not be easily classified into the defined haplogroups and total only 4% of the chromosomes.

HD chromosomes are almost exclusively (95%) haplogroup A (Figure 3B). In contrast, haplogroup A accounts for only 53% of chromosomes from the general population (<27 CAG). Haplogroup C was also very common on control chromosomes (41%) but completely absent from disease chromosomes. Similar to the HD chromosomes, 27–35 CAG chromosomes are enriched (83%) for haplogroup A relative to controls. It is also notable that each individual from the general population (<27 CAG) could be phased for high CAG and low CAG within the normal range, the higher CAG (mean CAG = 19.8 ± 2.7) chromosome was also statistically enriched (chi-square $p = 0.041$) for haplogroup A (62%) relative to the low CAG (mean CAG = 17.3 ± 2.1) chromosome (43%).

The CAG sizes for all chromosomes from 10–50 CAG-containing haplogroup A or C are plotted in Figure 3C. For the chromosomes used in this study, the mean CAG size for haplogroup A (33.9 ± 11.7) is significantly ($p < 0.00001$, t test) greater than that for haplogroup C (18.9 ± 5.9).

Although haplogroup A and C are both found on control chromosomes (<27 CAG), haplogroup A is uniquely enriched on chromosomes with an expanded CAG tract. Even when considering only control chromosomes (<27 CAG), the mean CAG of haplogroup A (18.8 ± 3.0) has a small but significant increase (t test $p < 0.00001$) compared to haplogroup C (mean = 17.5 ± 2.3). The odds ratio of haplogroup A (odds ratio 8.4, 95% CI from 4.6–15.6, chi-square $p = 5.9 \times 10^{-14}$) indicates that chromosomes >26 CAG are 8.4-fold more likely to occur on haplogroup A than other haplogroups. In comparison, the low odds ratio of haplogroup C (odds ratio 0.1, 95% CI from 0.03–0.17, chi-square $p = 1.2 \times 10^{-13}$) indicates that this haplogroup is significantly protected from CAG expansion.

Variants of Haplogroup A on CAG-Expanded Chromosomes

Haplogroup A is present on almost all CAG-expanded chromosomes but only ~50% of control chromosomes. To determine whether there were differences between haplogroup A when found on disease and control chromosomes, haplogroup A was subdivided into variants by subtracting the core elements that define haplogroup A and assessing the tSNPs that remain. Haplogroup A was defined by 10 tSNPs (3, 11, 14, 22, 65, 89, 97, 119, 169, and 178), so the haplogroup variants are defined by tSNPs at the remaining positions (tSNP 1, 55, 63, 69, 80, 95, 112, 176, 181, 182, 185, and 190). Haplogroup variants A1–A5 capture 98% of all haplogroup A chromosomes (Figure 4A). The remaining 2% of haplogroup A chromosomes were classified as "other," because they were singletons difficult to classify into variant groupings.

Of all of the HD chromosomes on haplogroup A, the majority (55%) can be classified as variant A1 (Figure 4A). Chromosomes with 27–35 CAG are also enriched for variant A1 (53%). In contrast, variants A4 and A5 are almost absent from expanded CAG chromosomes. Control chromosomes from the general population (<27 CAG) have a more even mixture of variants A1–A5. It is notable that in the general population (<27 CAG), variant A1 occurs more than three times more frequently on chromosomes with high-normal CAG versus low-normal CAG (Figure 4B).

The distribution of CAG sizes for all chromosomes carrying each haplogroup A variant is plotted in Figure 4C. Variant A1 occurs on chromosomes that range from 12 to 50 CAG, with the distribution shifted toward CAG expansion (mean CAG = 38.8 ± 9.5). Variant A2 also occurs on a range of CAG sizes extending from low normal (CAG 15) to high expanded (CAG 49) with an upward shift in mean CAG size (35 ± 11). Variant A3 has a bimodal distribution around normal and CAG-expanded chromosomes (mean CAG = 28 ± 12). Variants A4 (mean CAG = 20 ± 8) and A5 (mean CAG = 17 ± 5) are predominantly found on chromosomes with <27 CAG. The mean CAG is significantly different in all variants, with the exception of A4 versus A5 ($p < 0.001$, one-way ANOVA, Tukey posthoc).

Variant A1 confers the greatest odds ratio of CAG expansion (i.e., chromosomes with variant A1 are 6.5 times more likely to carry a CAG expansion, 95% CI from 3.5–12.3, chi-square $p = 8.5 \times 10^{-10}$). Variant A2 chromosomes are almost equally likely to carry a normal or expanded CAG (odds ratio 1.1, 95% CI from 0.6 to 1.9, chi-square $p = 8.3 \times 10^{-1}$). Variant A3 is almost twice as likely to contain a normal CAG versus expanded (odds ratio 0.5, 95% CI from 0.3 to 0.99, chi-square $p = 4.3 \times 10^{-2}$), whereas variants A4 and A5 are unlikely to carry a CAG expansion (odds ratio is close to 0, chi-square 95% CI from 0.00 to 0.25, $p = 1.2 \times 10^{-6}$ and 1.8×10^{-3}, respectively).

Loss of CAG tract interruption has been reported to confer increased CAG instability in *HTT*.[41–44] To determine whether the increased odds ratio of CAG expansion in A1 and A2 was due to loss of CAG tract interruption on these variants, direct sequencing was performed in exon 1 of *HTT*. Loss of interruption was not found in 10 HD and intermediate allele individuals sequenced, indicating no association between the risk haplotypes and loss of CAG tract interruption.

Taken together, these data suggest that there is an enrichment of specific haplotype variants on CAG-expanded chromosomes. Variants A1 and A2 confer the highest risk for having a CAG-expanded chromosome, whereas A4 and A5 variants are extensively protected from CAG expansion.

Haplogroups in the HapMap Cohort

We next wanted to determine whether the frequency of haplogroup A could account for differences in the prevalence of HD in ethnic groups in the HapMap populations. With the same haplogroup definitions (Figures 3 and 4), the frequency of haplogroups in each ethnic group is shown in Figure 5.

Interestingly, the frequency of haplogroup A is similar in Europe (46%), China (44%), and Japan (49%), even though the prevalence of HD in Asian populations (0.11 to 0.72 affected persons per 100,000[29,45–48]) is reported to be much lower than in Europe (5 to 7 per 100,000[2]). However, further analysis reveals that the Chinese and Japanese general population cohorts lack the presence of variants A1 and A2, the two variants with the highest odds ratio for CAG-expanded chromosomes. They also have a very high frequency of A5, a variant that is protected from CAG expansion.

African black populations were historically believed to have a low prevalence of HD.[49] Accordingly, the general population in Yoruba lacks the risk haplotypes and is greatly enriched for the protective A4 variant. Although recent studies suggest that the prevalence in South African blacks may be higher than originally reported,[50,51] the prevalence of HD has not been directly measured in the Yoruban people of Nigeria. The Yoruban population also has a very high proportion of "other" haplogroups, composed of nonmatching haplotypes, which reflects the greater genetic diversity in this population.

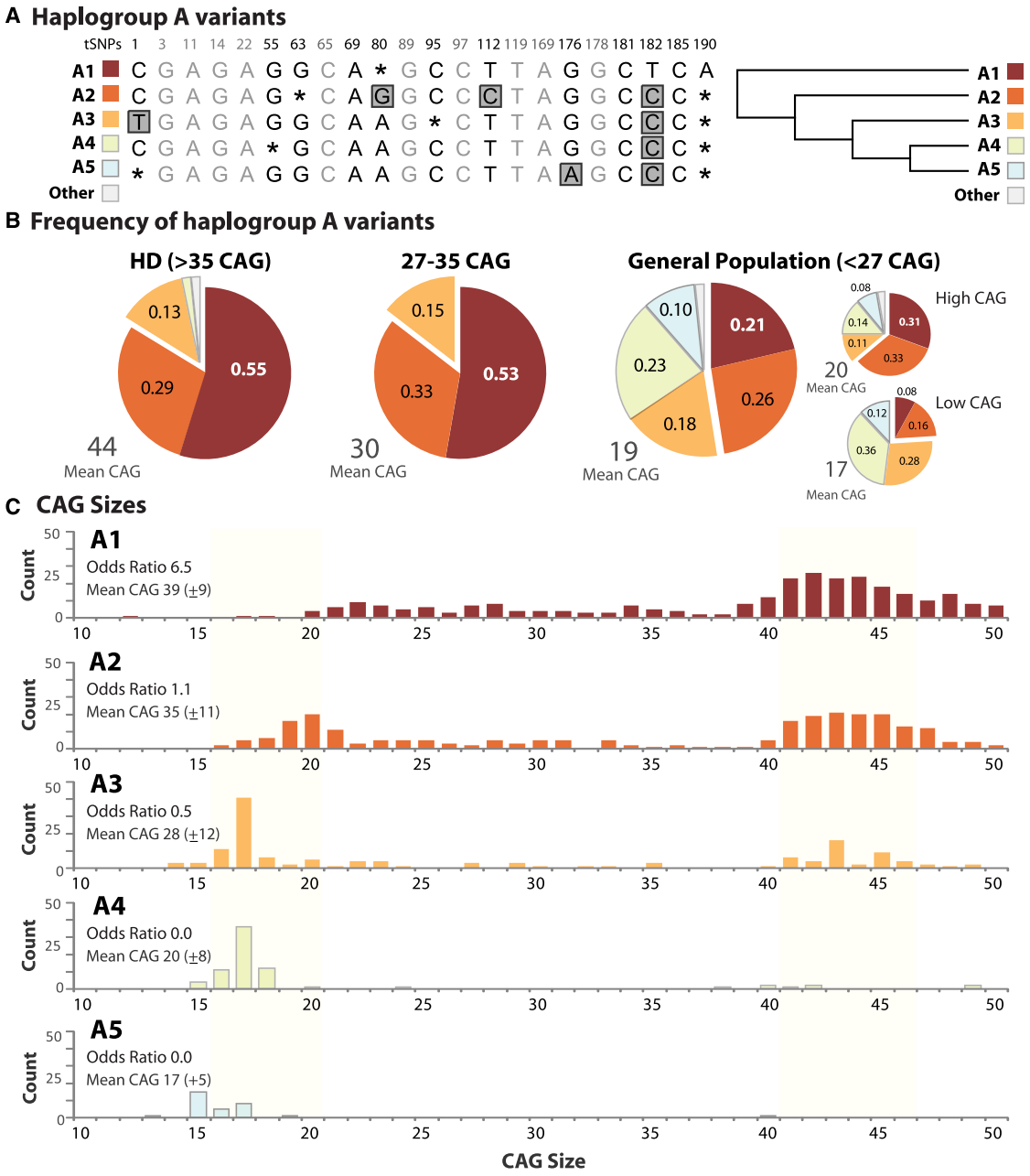

Figure 4. Specific Haplogroup A Variants Are Enriched on CAG-Expanded Chromosomes

(A) To determine whether there are differences in haplogroup A chromosomes from CAG-expanded and normal chromosomes, haplogroup A was divided into five major variants by subtracting the common tSNPs (light gray text) and with differences at the 12 remaining tSNP positions (black text). The wildcard asterisk (*) is used for variable allele positions. Dark gray boxes indicate differences relative to the A1 variant. The relationship between the variants is shown by a neighbor-joining tree (right).

(B) CAG-expanded chromosomes (HD, N = 62; 27–35 CAG carriers, n = 55) have similar haplogroup A variant distributions and are specifically enriched for A1 and A2 relative to chromosomes from the general population (n = 61). Phased chromosomes from the general population (right) demonstrates that large-normal chromosomes also have an enrichment for variant A1 and A2 relative to low-normal chromosomes. Variants A4 and A5 are almost absent from CAG-expanded chromosomes.

(C) CAG size distribution of chromosomes in each of subgroup. Variants A1, A2, and A3 chromosomes have a broad CAG size distribution that extends from low normal (<16 CAG) to high (>50). For the chromosomes used in this study, the mean CAG size and odds ratio of each variant is indicated. The highest HD risk variants, A1 and A2, have significantly elevated mean CAG size and odds ratios >1. Variant A3 is a moderate HD risk haplotype, because it has a larger component of CAG sizes in the normal range and therefore a lower mean CAG size. Chromosomes with variant A4 or A5 are stable in the normal range.

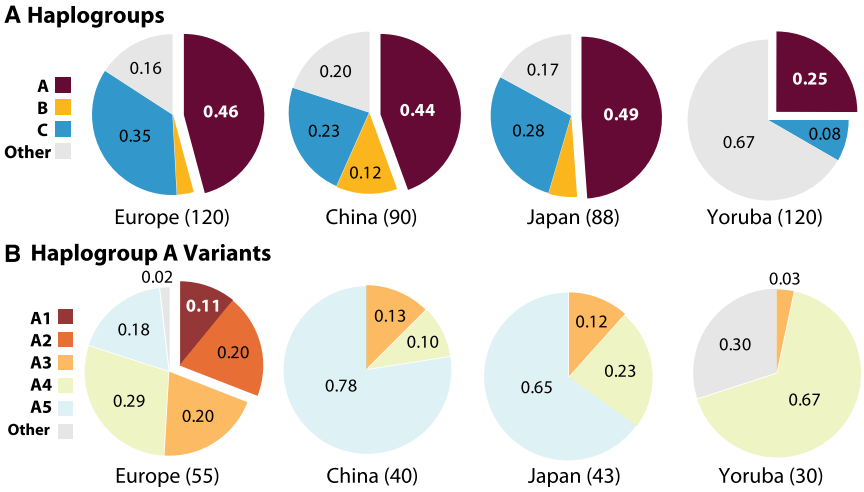

Figure 5. Ethnic Groups that Have a Low Prevalence of HD Do Not Have HD Risk Haplotypes in Their General Population

The prevalence of HD is much higher in Western European populations relative to Asia and Africa. Although the frequency of haplogroup A is similar between Europe and Asia (A), the frequencies of the high-risk variants of haplotype A, A1 and A2, are not found in the Asian populations (B). As expected, there is more genetic diversity in the Yoruba population, with a lower level of risk haplotypes and a relatively greater frequency of "other" haplotypes. Number of chromosomes assessed in each ethnic group is indicated in brackets.

The frequency of the haplogroup A variants appears to be significantly associated with differences in HD prevalence. Risk haplogroup variants for CAG expansion (A1 and A2) are absent and protected haplogroup variants (A4 and A5) are much more frequent in ethnic populations with a low prevalence of HD.

Identifying Targets for Allele-Specific Silencing of Mutant htt

Gene silencing has great potential for the treatment of HD and other trinucleotide repeat disorders.[52] To facilitate allele-specific gene silencing technologies for HD patients, we sought to determine whether the pattern of disease-associated alleles would allow the efficient and specific targeting of the mutant chromosome in the HD patient population. Because of the considerable expense involved in the clinical development of each target, it is important to be able to maximize the potential coverage of the HD population with a minimum number of allele targets (by using antisense oligonucleotides for gene silencing). A patient would theoretically be eligible for allele-specific treatment with a single targeting oligonucleotide if genotyping determined that they were heterozygous at the target SNP with the correct target allele associated with the disease chromosome. A crucial question is: Which target (or set of targets) would most efficiently cover the HD patient population and therefore be appropriate choices for drug development?

We report the coverage of the HD patient cohort in this study by using target panels comprising between 1 and 4 SNP targets (Figure 6). Multiple SNP combinations constitute each panel (Figure S5). Patient "coverage" is defined as the percent of the population of known genotypes that would be eligible for treatment. Seven of the 65 HD patients were not heterozygous at any SNP position, and therefore the maximum theoretical coverage is 89% (58/65). The maximum coverage by any single SNP was 52% (tSNP 182). Panels including 4 or 5 SNPs provided only a small increase in the coverage compared to the 3 SNP panel, which targeted all but 3% of the theoretically targetable HD patients in the cohort. In addition to the tSNP targets reported here, further SNPs in LD with these SNPs will also serve as equally suitable targets, providing further flexibility to oligonucleotide design. Many technical challenges remain pertaining to oligonucleotide design for efficient and specific silencing, but the presence of SNPs strongly associated with disease chromosomes provide an attractive opportunity for the development of allele-specific targeting.

Discussion

In this comprehensive analysis of the *HTT* region, we validated 190 HapMap SNPs and genotyped 22 nonredundant

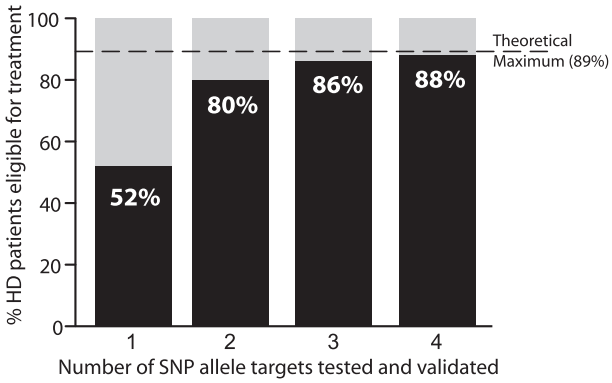

Figure 6. Disease-Associated SNPs Can Be Efficiently Targeted for Allele-Specific Silencing of the Mutant *HTT* mRNA

In an HD patient whose genotype is known, specific heterozygous alleles can be used to distinguish the CAG-expanded mRNA from nonexpanded mRNA (i.e., alleles that are 100% sensitive of the disease allele and 100% specific). Because of the expense of clinically testing and validating each target, it is important to maximize the patient coverage with a minimum number of targets. A theoretical maximum number of targetable patients (89%) exists because in this cohort, 7 of the 65 HD patients were not heterozygous at any tSNP and therefore could not be targeted. The maximum percent of the HD population in this study that could be treating with a single target (disease-associated allele) is 52%. The development of a therapy toward a second allele target would increase the patient coverage to 80%.

tagging SNPs from HD (>36 CAG), intermediate allele (27–35 CAG), and control (<27 CAG) chromosomes from individuals of European origin. We find the distribution of genotypes to be significantly different between CAG-expanded chromosomes (>26 CAG) and control chromosomes in the normal CAG range (<27 CAG). The disease-associated SNPs constitute a cluster of similar haplotypes (haplogroup A) found on 95% of disease chromosomes (odds ratio = 8.4, chi-square $p = 5.9 \times 10^{-14}$).

The association of many SNPs with CAG-expanded chromosomes is much stronger than previously reported polymorphisms in *HTT*[15–30] and are confirmed in a replication cohort. In addition, we now show that many SNPs are sensitive markers of disease chromosomes, meaning that a single SNP allele was almost always associated with CAG-expanded chromosomes. Importantly, two variants of haplogroup A (A1 and A2) were dramatically and specifically enriched on HD chromosomes and are therefore at increased risk for CAG expansion (odds ratio 6.5 and 1.1, respectively).

Interestingly, the same set of sensitive allelic markers and associated SNPs are also observed on chromosomes from 27–35 CAG carriers who are unaffected by HD but may be at risk of transmitting a chromosome that has expanded into the affected CAG range. Similar to disease chromosomes, haplogroup A was found on the majority (83%) of 27–35 CAG chromosomes.

There were no significant differences in SNP association between HD and 27–35 CAG chromosomes, suggesting that de novo mutations for HD continue to occur on the same haplogroup background as the rest of the HD population. These data suggest that chromosomes with CAG expansions above the normal range (>26 CAG) arise on a few haplotype variants that contain common elements that confer predisposition to expansion.

Origins of HD Mutations

The prevalence of HD is influenced by the balance between CAG-expanded chromosomes that are eliminated from the population because they are not transmitted to the next generation (i.e., juvenile HD) and new mutations occurring in families without a history of HD. The level of instability is greater for longer uninterrupted CAG repeat tracts and these are biased toward expansion.[53] What factor(s) determine which chromosomes will be unstable and undergo progressive CAG expansion over many generations? We propose three basic population models of CAG expansion in *HTT*.

Model I: Stochastic Expansion

If CAG expansion occurs randomly, then chromosomes with an equal CAG tract size should have an equal probability of expansion. Therefore, this stochastic model, where local DNA sequence has little influence on the CAG instability, would predict that the diversity of haplotypes in *HTT* would be similar in both HD patients and the general population (Figure 7, I).

This strictly stochastic model is clearly not supported by the current data. The vast majority of CAG-expanded chromosomes in the European population are found only on specific haplogroup variants (A1 and A2). Both longer-normal (20–26 CAG) and intermediate allele (27–35 CAG) chromosomes are specifically enriched for the same haplogroup as the HD patients. Notably, there are haplogroups (C) and haplogroup variants (A4 and A5) on which CAG expansion almost never occurs. CAG expansion occurs preferentially on specific haplotypes and is therefore not a stochastic process.

Similarly, if genetic *trans*-factors or environmental influences are the primary determinants of CAG instability in *HTT* (Figure 7, I), we would not expect a dramatic enrichment of a specific *HTT* haplogroup with CAG expansion. If factors such as mutations in mismatch repair genes, the age of parents, or environmental toxins alone were sufficient to cause CAG instability, we would not expect the majority of CAG-expanded chromosomes to be restricted to a limited number of haplotypes.

Model II: Founder Chromosome

The presence of HD chromosomes on a single haplogroup could be due to the coincidental linkage of markers with an initial (founder) CAG-expansion event. In this model, a single (or limited number) of chromosomes underwent a large expansion in CAG size, and all CAG-expanded chromosomes are derived from this founder (Figure 7, II). A variation of the founder hypothesis would be a rare mutational event that resulted in an intermediate or large-normal allele, producing a pool of chromosomes not affected by HD, but at risk for expansion in later generations because of their larger-than-normal CAG size, such as observed in Friedreich Ataxia (FRDA [MIM 229300])[54] and specific populations of Myotonic Dystrophy 1 (DM1 [MIM 160900]).[55–58]

The pattern of SNPs on HD chromosomes is not consistent with a founder CAG-expansion event for several reasons: (1) there are multiple haplogroup variants (A1 and A2) associated with CAG expansion that are not easily derived from one another; (2) the pattern of strong SNP linkage to CAG expansion is punctuated, rather than decaying as a function of genetic distance; and (3) there are no haplotypes found only on CAG-expanded chromosomes and not on control chromosomes, as would be expected from a stratified founder population.

(1) Unlike other trinucleotide repeat diseases believed to originate from a founder chromosome, CAG-expanded chromosomes in HD arise on multiple haplogroup variants. DM1, for example, is believed to arise from a founder chromosome with a large-normal CTG tract in the *DMPK* gene (MIM 605377), which serves as a reservoir for expanded CTG chromosomes because of its already large-normal CTG size.[55] This single haplotype is associated with CTG-expanded chromosomes at numerous consecutive markers.[57] French Canadian DM1 families, for example, were found to contain a core 500 kb region of identical sequence consistent with a single ancestral

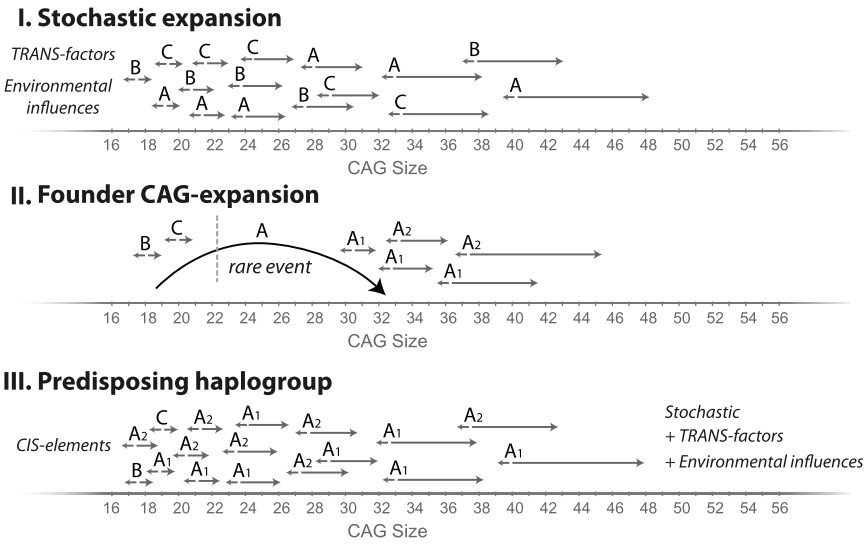

Figure 7. Pattern of Disease-Associated SNPs Is Consistent with a Model of Stepwise CAG Expansion on a Predisposing Haplogroup

Schematic of possible models of CAG expansion. Arrows represent mutational events that result in a change in CAG tract size across generations on specific chromosomal haplogroups.

(I) Stochastic model results in essentially unpredictable CAG expansion over many generations. If the primary factors influencing CAG instability are other than *cis*-elements (i.e., genetic *trans*-factors, environmental influences, or entirely stochastic processes), then CAG expansions should occur randomly on chromosomes with random genotypes. HD chromosomes should therefore have a distribution of haplogroups similar to the general population.

(II) The expanded CAG founder (or large normal founder) model predicts a rare and large expansion in the CAG tract in a limited number of individuals. CAG instability then results because of the overall increase in CAG tract size. All HD individuals are related to the founder chromosome(s) and would be expected to have an identical haplogroup except for instances where recombination events have occurred. The founder population is genetically isolated from the general population and unique haplotypes should arise on CAG-expanded chromosomes only.

(III) The predisposing haplogroup model predicts that CAG expansion occurs primarily on a haplogroup containing predisposing *cis*-elements. The *cis*-elements may interact with *trans*-factors or environmental influences, but the predisposing haplogroup (containing the *cis*-elements and polymorphisms in LD with the *cis*-elements) is required for instability to occur and is therefore the predominant haplogroup of disease chromosomes. The source of the predisposing haplogroup is not unique to intermediate-sized CAG alleles, so the origin of unstable CAG chromosomes is the general population carrying the predisposing *cis*-elements. The significant enrichment of HD on haplogroup A is most consistent with model III.

founder.[56] This is in contrast to European HD chromosomes, where the strength of allele associations is variable across the gene region. CAG expansion in HD arose multiple times on haplogroup variants that are not easily derived from one another but share elements in common.

(2) Although there are numerous SNPs (12/22) spanning the gene region that are both highly associated with and sensitive markers of CAG-expanded chromosomes, these are punctuated with other SNPs (6/22) that completely lack association with CAG expansion. If HD chromosomes were derived from a founder, they would be expected to have identical alleles at all SNP positions close to the mutation, with the strength of linkage decaying as a function of nucleotide distance and recent recombination events.[59] It is difficult to reconcile the lack of association and allele specificity interspersed across the gene (i.e., tSNP 63, 80, 112, and 182) with a recent single-founder hypothesis for HD.

(3) These data are not consistent with a founder or large-normal CAG founder for HD because the specific disease-associated haplogroup variants are also prevalent in the general population (25%). A large-normal CAG founder would be expected to result in a genetically isolated population in which unique haplotypes will arise. This is the case in FRDA, where the disease-associated haplotype from a long-normal founder is infrequent in the general population (4%).[54] In the European population, there are no haplotypes that were unique to CAG-expanded HD chromosomes.

Model III: Predisposing Haplogroup

The current data are most consistent with CAG expansion occurring on haplotypes that are predisposed for CAG instability (Figure 7, III). These haplotypes may contain DNA *cis*-elements that either introduce instability themselves or whose affinity for repeat-stabilizing factors has been compromised. The stability of other trinucleotide repeats in the genome is influenced by *cis*-elements,[60–62] including specific identified features such as the replication origin[63] and sequences that contribute to the formation of secondary structures.[64] Recently, it has been shown that mutation or methylation of the CTCF binding site adjacent to the CAG-tract of *ATXN7* (MIM 607640) promotes repeat instability in a mouse model of Spinocerebellar Ataxia 7 (SCA7 [MIM 164500]).[65]

A loss of CAG tract interruption (resulting in an extended pure CAG tract) is an example of another *cis*-element that plays an important role in instability in other CAG repeat diseases, such as Myotonic Dystrophy 2 (DM2 [MIM 602668]),[66] FRDA,[54] Fragile X (MIM 300624),[67] SCA1 (MIM 164400),[68] and SCA17 (MIM 607136).[69] However, this loss of interruption occurs only rarely in *HTT*[41–44] and is not associated with the haplogroup variants with the highest risk of CAG expansion.

It is not clear what *cis*-elements on haplogroup variants A1 and A2 are increasing CAG instability in *HTT*. The complementary hypothesis is that *cis*-elements on the

non-HD haplogroups (primarily haplogroup C and haplogroup A variants A4 and A5) provide stability to the CAG tract. Of particular interest are the tSNPs that distinguish haplogroup A from haplogroup C (tSNP 11, 14, and 89) and other unknown polymorphisms in LD with these. Because of the punctuated pattern of allele associations, multiple *cis*-elements may singly or cumulatively influence CAG stability.

Importantly, the haplogroup variants with the highest risk of CAG expansion (A1 and A2) are also found on general population chromosomes. This suggests that multistep expansions into the HD-affected range occur from a reservoir of normal alleles containing the predisposing *cis*-elements. These high-risk variants exist in a subset of the general population of Europe and allow prediction of which families may experience CAG expansion in future generations. It is plausible that the majority of European HD cases are due to initial *cis*-element mutations in *HTT* in the general population that increased CAG instability (rather than the founder mutation being the CAG expansion itself).

Although we can predict *which* chromosomes may be susceptible to CAG expansion, *when* CAG tracts expand appears to be more stochastic and could have many influences. In the presence of the predisposing haplotype, *trans*-factors (i.e., DNA repair genes[11]), environmental influences (sex and age of the transmitting parent, diet, toxins), and primarily the length of uninterrupted CAG repeats will influence the timing and level of CAG expansion on transmitted chromosomes.

Geographic Variability in HD

We find that haplogroup A variants with the highest odds ratio for CAG expansion (A1 and A2) are not present in ethnic populations with a low prevalence of HD (Figure 5). Epidemiological reports have established prevalence rates ranging from 5 to 10 affected persons per 100,000 in countries of Western European ancestry, including the United States and Canada.[70–73] Considerably reduced frequencies of HD have been reported in Japan,[29,45,46] China,[47,48] Finland,[74] South African black populations,[49–51] and North American black populations.[71] The presence of the predisposing haplogroup variants in the general population of Western Europe provides a compelling explanation for the >10-fold increase in prevalence relative to other ethnic groups.

In the few cases where expansions have occurred on low-risk haplogroup variants—both in the European and other ethnic groups—it is possible that a low background level of stochastic expansion (model I) may allow for CAG expansion in the absence of predisposing *cis*-elements. The molecular mechanism of CAG instability may be different between stochastic expansion and CAG expansion on a predisposing haplotype. Future studies will need to control for haplotypes when assessing the detailed mechanisms of CAG instability at the molecular level.

Gene Silencing

Several lines of evidence suggest that gene silencing of the mutant *HTT* could be a successful treatment for HD. Previous studies in mouse models of HD have demonstrated that the severity of the disease is dependent on the level of expression of the mutant gene.[75–82]

However, wild-type huntingtin has a crucial function during development of the nervous system and a protective role in cells.[83,84] Constitutive knockout of *HTT* is lethal during embryonic development[85–87] and adult inactivation of *HTT* leads to progressive degeneration in the brain and the testes.[88] In addition, because of aggregation, the half life of mutant htt may be longer than WT htt, and non-allele-specific silencing might increase the ratio of mutant to wild-type htt and worsen the disease phenotype. Allele-specific silencing of the mutant gene would be desirable to avoid the potentially negative consequences of reducing huntingtin expression from the wild-type allele.

Precise silencing of genes or nucleic-acid gene products can be achieved with sequence-based targeting methods, such as RNA interference (RNAi) technology, antisense oligonucleotides (ASO), microRNA (miRNA), or small hairpin interference RNA (shRNA). Unfortunately, the CAG tract expansion that causes HD cannot be targeted directly because of its size and repetitive nature. Instead, SNPs that are in LD with expanded CAG alleles would be attractive surrogate targets for allele-specific gene silencing of mutant *HTT*.

The high degree of linkage observed between disease chromosomes and specific alleles provides the therapeutic opportunity to selectively knock down the disease allele. Allele-specific silencing of mutant *HTT* has been recently achieved in patient-derived fibroblasts with RNAi.[89,90]

It now becomes important to identify which disease-associated SNP alleles most efficiently target the HD patient population. The expense of drug development for each target is considerable, and it is therefore important to maximize patient population coverage with a minimum number of therapeutic targets. With this information on efficient targets, the challenge becomes the development and testing of oligonucleotides that effectively and specifically silence the mutant allele. The information provided here (Figure 6; Figure S5) will enable multiple technologies for this purpose.

Conclusions

CAG expansion in *HTT* is strongly associated with a set of SNPs that constitute a single haplogroup. The data are consistent with a multistep model for CAG expansion in HD and provides evidence that CAG instability is most likely to occur on predisposing haplotypes containing specific *cis*-elements. The high prevalence of HD in Europeans relative to other populations may be due to an initial *cis*-element mutation in the European general population, which conferred a greater probability of CAG instability. Because of this, we hypothesize that there exists a subpopulation of unaffected carriers (both intermediate allele and

in the general population) that have a predisposition to CAG expansion and may result in new mutations for HD in future generations.

These data are not consistent with the HD mutation occurring randomly or primarily as a result of genetic *trans*-factors or environmental influences because CAG expansion occurs on specific, nonrandom haplogroup variants. The pattern of allele associations and the population distribution of these at-risk haplogroup variants make them unlikely to be the result of a founder CAG expansion for the origins of HD.

Many of the SNPs highly associated with CAG expansion do not segregate independently and are in LD with each other. Future studies will seek to identify which SNPs may point to *cis*-elements for CAG instability in *HTT*. Finally, the strong LD between a specific haplogroup and CAG-expanded chromosomes provides an opportunity for the therapeutic development of personalized medicine in HD. The efficient allele-specific targeting of mutant huntingtin in the majority of HD patients may require the clinical validation of only a small number of targeting oligonucleotides.

Supplemental Data

Supplemental Data include five figures and can be found with this article online at http://www.ajhg.org/.

Acknowledgments

The authors thank HD and intermediate allele families for the generous use of their DNA for this research. We also thank Colin Ross, Hagit Katzov, Terry Pape, Rebecca Devon, Fiona Young, and Mahmoud Pouladi for helpful discussions and critical reading of the manuscript. Funding for these studies has been provided by the Canadian Genetic Diseases Network (CGDN), Huntington Society of Canada (HSC), Michael Smith Foundation for Health Research (MSFHR), Canadian Institutes of Health Research (CIHR), Child and Family Research Institute (CFRI), and the HDSA/High Q Foundation. J.B.C. is funded by MSFHR and CIHR. M.R.H. is a Killam University Professor and holds a Canada Research Chair in Human Genetics and Molecular Medicine. The authors report no conflicts of interest associated with this work.

Received: November 3, 2008
Revised: January 13, 2009
Accepted: February 3, 2009
Published online: February 26, 2009

Web Resources

The URLs for data presented herein are as follows:

Autoprimer software, http://www.autoprimer.com
dbSNP, http://www.ncbi.nlm.nih.gov/SNP
International HapMap Project, http://www.hapmap.org
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/
SNP Consortium, http://snp.cshl.org

References

1. Kremer, B., Goldberg, P., Andrew, S.E., Theilmann, J., Telenius, H., Zeisler, J., Squitieri, F., Lin, B., Bassett, A., Almqvist, E., et al. (1994). A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. N. Engl. J. Med. *330*, 1401–1406.
2. Walker, F.O. (2007). Huntington's disease. Lancet *369*, 218–228.
3. Telenius, H., Kremer, H.P., Theilmann, J., Andrew, S.E., Almqvist, E., Anvret, M., Greenberg, C., Greenberg, J., Lucotte, G., Squitieri, F., et al. (1993). Molecular analysis of juvenile Huntington disease: the major influence on (CAG)n repeat length is the sex of the affected parent. Hum. Mol. Genet. *2*, 1535–1540.
4. Falush, D., Almqvist, E.W., Brinkmann, R.R., Iwasa, Y., and Hayden, M.R. (2000). Measurement of mutational flow implies both a high new-mutation rate for Huntington disease and substantial underascertainment of late-onset cases. Am. J. Hum. Genet. *68*, 373–385.
5. Almqvist, E.W., Elterman, D.S., MacLeod, P.M., and Hayden, M.R. (2001). High incidence rate and absent family histories in one quarter of patients newly diagnosed with Huntington disease in British Columbia. Clin. Genet. *60*, 198–205.
6. Nguyen, G.H., Bouchard, J., Boselli, M.G., Tolstoi, L.G., Keith, L., Baldwin, C., Nguyen, N.C., Schultz, M., Herrera, V.L., and Smith, C.L. (2003). DNA stability and schizophrenia in twins. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *120B*, 1–10.
7. Norremolle, A., Hasholt, L., Petersen, C.B., Eiberg, H., Hasselbalch, S.G., Gideon, P., Nielsen, J.E., and Sorensen, S.A. (2004). Mosaicism of the CAG repeat sequence in the Huntington disease gene in a pair of monozygotic twins. Am. J. Med. Genet. A. *130A*, 154–159.
8. Semaka, A., Creighton, S., Warby, S., and Hayden, M.R. (2006). Predictive testing for Huntington disease: interpretation and significance of intermediate alleles. Clin. Genet. *70*, 283–294.
9. Pearson, C.E., Edamura, K.N., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. *6*, 729–742.
10. McMurray, C.T. (2008). Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. DNA Repair (Amst.) *7*, 1121–1134.
11. Manley, K., Shirley, T.L., Flaherty, L., and Messer, A. (1999). Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat. Genet. *23*, 471–473.
12. Kovtun, I.V., Liu, Y., Bjoras, M., Klungland, A., Wilson, S.H., and McMurray, C.T. (2007). OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. Nature *447*, 447–452.
13. Cleary, J.D., and Pearson, C.E. (2003). The contribution of cis-elements to disease-associated repeat instability: clinical and experimental evidence. Cytogenet. Genome Res. *100*, 25–55.
14. Giovannone, B., Sabbadini, G., DiMaio, L., Calabrese, O., Castaldo, I., Frontali, M., Novelletto, A., and Squitieri, F. (1997). Analysis of (CAG)n size heterogeneity in somatic and sperm cell DNA from intermediate and expanded Huntington disease gene carriers. Hum. Mutat. *10*, 458–464.
15. Rubinsztein, D.C., Leggo, J., Goodburn, S., Barton, D.E., and Ferguson-Smith, M.A. (1995). Haplotype analysis of the delta 2642 and (CAG)n polymorphisms in the Huntington's disease

(HD) gene provides an explanation for an apparent 'founder' HD haplotype. Hum. Mol. Genet. *4*, 203–206.
16. Scholefield, J., and Greenberg, J. (2007). A common SNP haplotype provides molecular proof of a founder effect of Huntington disease linking two South African populations. Eur. J. Hum. Genet. *15*, 590–595.
17. Kartsaki, E., Spanaki, C., Tzagournissakis, M., Petsakou, A., Moschonas, N., MacDonald, M., and Plaitakis, A. (2006). Late-onset and typical Huntington disease families from Crete have distinct genetic origins. Int. J. Mol. Med. *17*, 335–346.
18. Garcia-Planells, J., Burguera, J.A., Solis, P., Millan, J.M., Ginestar, D., Palau, F., and Espinos, C. (2005). Ancient origin of the CAG expansion causing Huntington disease in a Spanish population. Hum. Mutat. *25*, 453–459.
19. Wang, C.K., Wu, Y.R., Hwu, W.L., Chen, C.M., Ro, L.S., Chen, S.T., Gwinn-Hardy, K., Yang, C.C., Wu, R.M., Chen, T.F., et al. (2004). DNA haplotype analysis of CAG repeat in Taiwanese Huntington's disease patients. Eur. Neurol. *52*, 96–100.
20. Saleem, Q., Roy, S., Murgood, U., Saxena, R., Verma, I.C., Anand, A., Muthane, U., Jain, S., and Brahmachari, S.K. (2003). Molecular analysis of Huntington's disease and linked polymorphisms in the Indian population. Acta Neurol. Scand. *108*, 281–286.
21. Pramanik, S., Basu, P., Gangopadhaya, P.K., Sinha, K.K., Jha, D.K., Sinha, S., Das, S.K., Maity, B.K., Mukherjee, S.C., Roychoudhuri, S., et al. (2000). Analysis of CAG and CCG repeats in Huntingtin gene among HD patients and normal populations of India. Eur. J. Hum. Genet. *8*, 678–682.
22. Hecimovic, S., Klepac, N., Vlasic, J., Vojta, A., Janko, D., Skarpa-Prpic, I., Canki-Klain, N., Markovic, D., Bozikov, J., Relja, M., et al. (2002). Genetic background of Huntington disease in Croatia: molecular analysis of CAG, CCG, and Delta2642 (E2642del) polymorphisms. Hum. Mutat. *20*, 233.
23. Dode, C., Durr, A., Pecheux, C., Mouret, J.F., Belal, S., Bachner, L., Agid, Y., Kaplan, J.C., Brice, A., and Feingold, J. (1993). Huntington's disease in French families: CAG repeat expansion and linkage disequilibrium analysis. C. R. Acad. Sci. III *316*, 1374–1380.
24. Paradisi, I., Hernandez, A., and Arias, S. (2008). Huntington disease mutation in Venezuela: age of onset, haplotype analyses and geographic aggregation. J. Hum. Genet. *53*, 127–135.
25. Almqvist, E., Andrew, S., Theilmann, J., Goldberg, P., Zeisler, J., Drugge, U., Grandell, U., Tapper-Persson, M., Winblad, B., Hayden, M., et al. (1994). Geographical distribution of haplotypes in Swedish families with Huntington's disease. Hum. Genet. *94*, 124–128.
26. Andrew, S., Theilmann, J., Almqvist, E., Norremolle, A., Lucotte, G., Anvret, M., Sorensen, S.A., Turpin, J.C., and Hayden, M.R. (1993). DNA analysis of distinct populations suggests multiple origins for the mutation causing Huntington disease. Clin. Genet. *43*, 286–294.
27. Squitieri, F., Andrew, S.E., Goldberg, Y.P., Kremer, B., Spence, N., Zeisler, J., Nichol, K., Theilmann, J., Greenberg, J., Goto, J., et al. (1994). DNA haplotype analysis of Huntington disease reveals clues to the origins and mechanisms of CAG expansion and reasons for geographic variations of prevalence. Hum. Mol. Genet. *3*, 2103–2114.
28. Costa, M.C., Magalhaes, P., Guimaraes, L., Maciel, P., Sequeiros, J., and Sousa, A. (2006). The CAG repeat at the Huntington disease gene in the Portuguese population: insights into its dynamics and to the origin of the mutation. J. Hum. Genet. *51*, 189–195.
29. Masuda, N., Goto, J., Murayama, N., Watanabe, M., Kondo, I., and Kanazawa, I. (1995). Analysis of triplet repeats in the huntingtin gene in Japanese families affected with Huntington's disease. J. Med. Genet. *32*, 701–705.
30. Almqvist, E., Spence, N., Nichol, K., Andrew, S.E., Vesa, J., Peltonen, L., Anvret, M., Goto, J., Kanazawa, I., Goldberg, Y.P., et al. (1995). Ancestral differences in the distribution of the delta 2642 glutamic acid polymorphism is associated with varying CAG repeat lengths on normal chromosomes: insights into the genetic evolution of Huntington disease. Hum. Mol. Genet. *4*, 207–214.
31. Andrew, S.E., and Hayden, M.R. (1995). Origins and evolution of Huntington disease chromosomes. Neurodegeneration *4*, 239–244.
32. International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
33. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
34. Kumar, S., Tamura, K., and Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. *5*, 150–163.
35. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263–265.
36. de Bakker, P.I., Graham, R.R., Altshuler, D., Henderson, B.E., and Haiman, C.A. (2006). Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. Pac. Symp. Biocomput. 478–486.
37. Bell, P.A., Chaturvedi, S., Gelfand, C.A., Huang, C.Y., Kochersperger, M., Kopla, R., Modica, F., Pohl, M., Varde, S., Zhao, R., et al. (2002). SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. Biotechniques (*Suppl*), 70–77.
38. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., Defelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225–2229.
39. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072–1079.
40. Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M., Cardon, L., Hudson, T.J., and Metspalu, A. (2006). An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. PLoS. Genet. *2*, e27.
41. Goldberg, Y.P., McMurray, C.T., Zeisler, J., Almqvist, E., Sillence, D., Richards, F., Gacy, A.M., Buchanan, J., Telenius, H., and Hayden, M.R. (1995). Increased instability of intermediate alleles in families with sporadic Huntington disease compared to similar sized intermediate alleles in the general population. Hum. Mol. Genet. *4*, 1911–1918.
42. Gellera, C., Meoni, C., Castellotti, B., Zappacosta, B., Girotti, F., Taroni, F., and Didonato, S. (1996). Errors in Huntington disease diagnostic test caused by trinucleotide deletion in the IT15 gene. Am. J. Hum. Genet. *59*, 475–477.
43. Williams, L.C., Hegde, M.R., Nagappan, R., Faull, R.L., Giles, J., Winship, I., Snow, K., and Love, D.R. (2000). Null alleles at the Huntington disease locus: implications for diagnostics and CAG repeat instability. Genet. Test. *4*, 55–60.

44. Yu, S., Fimmel, A., Fung, D., and Trent, R.J. (2000). Polymorphisms in the CAG repeat—a source of error in Huntington disease DNA testing. Clin. Genet. *58*, 469–472.
45. Nakashima, K., Watanabe, Y., Kusumi, M., Nanba, E., Maeoka, Y., Nakagawa, M., Igo, M., Irie, H., Ishino, H., Fujimoto, A., et al. (1996). Epidemiological and genetic studies of Huntington's disease in the San-in area of Japan. Neuroepidemiology *15*, 126–131.
46. Adachi, Y., and Nakashima, K. (1999). [Population genetic study of Huntington's disease—prevalence and founder's effect in the San-in area, western Japan.] Nippon Rinsho *57*, 900–904.
47. Leung, C.M., Chan, Y.W., Chang, C.M., Yu, Y.L., and Chen, C.N. (1992). Huntington's disease in Chinese: a hypothesis of its origin. J. Neurol. Neurosurg. Psychiatry *55*, 681–684.
48. Chang, C.M., Yu, Y.L., Fong, K.Y., Wong, M.T., Chan, Y.W., Ng, T.H., Leung, C.M., and Chan, V. (1994). Huntington's disease in Hong Kong Chinese: epidemiology and clinical picture. Clin. Exp. Neurol. *31*, 43–51.
49. Hayden, M.R., MacGregor, J.M., and Beighton, P.H. (1980). The prevalence of Huntington's chorea in South Africa. S. Afr. Med. J. *58*, 193–196.
50. Magazi, D.S., Krause, A., Bonev, V., Moagi, M., Iqbal, Z., Dludla, M., and van der Meyden, C.H. (2008). Huntington's disease: genetic heterogeneity in black African patients. S. Afr. Med. J. *98*, 200–203.
51. Silber, E., Kromberg, J., Temlett, J.A., Krause, A., and Saffer, D. (1998). Huntington's disease confirmed by genetic testing in five African families. Mov. Disord. *13*, 726–730.
52. Bonini, N.M., and La Spada, A.R. (2005). Silencing polyglutamine degeneration with RNAi. Neuron *48*, 715–718.
53. Rubinsztein, D.C., Amos, W., Leggo, J., Goodburn, S., Ramesar, R.S., Old, J., Bontrop, R., McMahon, R., Barton, D.E., and Ferguson-Smith, M.A. (1994). Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. Nat. Genet. *7*, 525–530.
54. Cossee, M., Schmitt, M., Campuzano, V., Reutenauer, L., Moutou, C., Mandel, J.L., and Koenig, M. (1997). Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations. Proc. Natl. Acad. Sci. USA *94*, 7452–7457.
55. Imbert, G., Kretz, C., Johnson, K., and Mandel, J.L. (1993). Origin of the expansion mutation in myotonic dystrophy. Nat. Genet. *4*, 72–76.
56. Whiting, E.J., Tsilfidis, C., Surh, L., MacKenzie, A.E., and Korneluk, R.G. (1995). Convergent myotonic dystrophy (DM) haplotypes: potential inconsistencies in human disease gene localization. Eur. J. Hum. Genet. *3*, 195–202.
57. Neville, C.E., Mahadevan, M.S., Barcelo, J.M., and Korneluk, R.G. (1994). High resolution genetic analysis suggests one ancestral predisposing haplotype for the origin of the myotonic dystrophy mutation. Hum. Mol. Genet. *3*, 45–51.
58. Tishkoff, S.A., Goldman, A., Calafell, F., Speed, W.C., Deinard, A.S., Bonne-Tamir, B., Kidd, J.R., Pakstis, A.J., Jenkins, T., and Kidd, K.K. (1998). A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. Am. J. Hum. Genet. *62*, 1389–1402.
59. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.
60. La Spada, A.R., Peterson, K.R., Meadows, S.A., McClain, M.E., Jeng, G., Chmelar, R.S., Haugen, H.A., Chen, K., Singer, M.J., Moore, D., et al. (1998). Androgen receptor YAC transgenic mice carrying CAG 45 alleles show trinucleotide repeat instability. Hum. Mol. Genet. *7*, 959–967.
61. Libby, R.T., Monckton, D.G., Fu, Y.H., Martinez, R.A., McAbney, J.P., Lau, R., Einum, D.D., Nichol, K., Ware, C.B., Ptacek, L.J., et al. (2003). Genomic context drives SCA7 CAG repeat instability, while expressed SCA7 cDNAs are intergenerationally and somatically stable in transgenic mice. Hum. Mol. Genet. *12*, 41–50.
62. Ennis, S., Murray, A., Brightwell, G., Morton, N.E., and Jacobs, P.A. (2007). Closely linked cis-acting modifier of expansion of the CGG repeat in high risk FMR1 haplotypes. Hum. Mutat. *28*, 1216–1224.
63. Cleary, J.D., Nichol, K., Wang, Y.H., and Pearson, C.E. (2002). Evidence of cis-acting factors in replication-mediated trinucleotide repeat instability in primate cells. Nat. Genet. *31*, 37–46.
64. Moore, H., Greenwell, P.W., Liu, C.P., Arnheim, N., and Petes, T.D. (1999). Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. USA *96*, 1504–1509.
65. Libby, R.T., Hagerman, K.A., Pineda, V.V., Lau, R., Cho, D.H., Baccam, S.L., Axford, M.M., Cleary, J.D., Moore, J.M., Sopher, B.L., et al. (2008). CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner: a novel basis for mutational hot spot determination. PLoS. Genet. *4*, e1000257.
66. Liquori, C.L., Ikeda, Y., Weatherspoon, M., Ricker, K., Schoser, B.G., Dalton, J.C., Day, J.W., and Ranum, L.P. (2003). Myotonic dystrophy type 2: human founder haplotype and evolutionary conservation of the repeat tract. Am. J. Hum. Genet. *73*, 849–862.
67. Eichler, E.E., Holden, J.J., Popovich, B.W., Reiss, A.L., Snow, K., Thibodeau, S.N., Richards, C.S., Ward, P.A., and Nelson, D.L. (1994). Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Nat. Genet. *8*, 88–94.
68. Mulvihill, D.J., Nichol, E.K., Hagerman, K.A., Pearson, C.E., and Wang, Y.H. (2005). Effect of CAT or AGG interruptions and CpG methylation on nucleosome assembly upon trinucleotide repeats on spinocerebellar ataxia, type 1 and fragile X syndrome. J. Biol. Chem. *280*, 4498–4503.
69. Gao, R., Matsuura, T., Coolbaugh, M., Zuhlke, C., Nakamura, K., Rasmussen, A., Siciliano, M.J., Ashizawa, T., and Lin, X. (2008). Instability of expanded CAG/CAA repeats in spinocerebellar ataxia type 17. Eur. J. Hum. Genet. *16*, 215–222.
70. Beighton, P., and Hayden, M.R. (1981). Huntington's chorea. S. Afr. Med. J. *59*, 250.
71. Folstein, S.E., Chase, G.A., Wahl, W.E., McDonnell, A.M., and Folstein, M.F. (1987). Huntington disease in Maryland: clinical aspects of racial variation. Am. J. Hum. Genet. *41*, 168–179.
72. Morrison, P.J., Johnston, W.P., and Nevin, N.C. (1995). The epidemiology of Huntington's disease in Northern Ireland. J. Med. Genet. *32*, 524–530.
73. Harper, P.S., Lim, C., and Craufurd, D. (2000). Ten years of presymptomatic testing for Huntington's disease: the experience of the UK Huntington's Disease Prediction Consortium. J. Med. Genet. *37*, 567–571.

74. Palo, J., Somer, H., Ikonen, E., Karila, L., and Peltonen, L. (1987). Low prevalence of Huntington's disease in Finland. Lancet *2*, 805–806.
75. Graham, R.K., Slow, E.J., Deng, Y., Bissada, N., Lu, G., Pearson, J., Shehadeh, J., Leavitt, B.R., Raymond, L.A., and Hayden, M.R. (2006). Levels of mutant huntingtin influence the phenotypic severity of Huntington disease in YAC128 mouse models. Neurobiol. Dis. *21*, 444–455.
76. Yamamoto, A., Lucas, J.J., and Hen, R. (2000). Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. Cell *101*, 57–66.
77. DiFiglia, M., Sena-Esteves, M., Chase, K., Sapp, E., Pfister, E., Sass, M., Yoder, J., Reeves, P., Pandey, R.K., Rajeev, K.G., et al. (2007). Therapeutic silencing of mutant huntingtin with siRNA attenuates striatal and cortical neuropathology and behavioral deficits. Proc. Natl. Acad. Sci. USA *104*, 17204–17209.
78. Machida, Y., Okada, T., Kurosawa, M., Oyama, F., Ozawa, K., and Nukina, N. (2006). rAAV-mediated shRNA ameliorated neuropathology in Huntington disease model mouse. Biochem. Biophys. Res. Commun. *343*, 190–197.
79. Wang, Y.L., Liu, W., Wada, E., Murata, M., Wada, K., and Kanazawa, I. (2005). Clinico-pathological rescue of a model mouse of Huntington's disease by siRNA. Neurosci. Res. *53*, 241–249.
80. Rodriguez-Lebron, E., Denovan-Wright, E.M., Nash, K., Lewin, A.S., and Mandel, R.J. (2005). Intrastriatal rAAV-mediated delivery of anti-huntingtin shRNAs induces partial reversal of disease progression in R6/1 Huntington's disease transgenic mice. Mol. Ther. *12*, 618–633.
81. Harper, S.Q., Staber, P.D., He, X., Eliason, S.L., Martins, I.H., Mao, Q., Yang, L., Kotin, R.M., Paulson, H.L., and Davidson, B.L. (2005). RNA interference improves motor and neuropathological abnormalities in a Huntington's disease mouse model. Proc. Natl. Acad. Sci. USA *102*, 5820–5825.
82. Franich, N.R., Fitzsimons, H.L., Fong, D.M., Klugmann, M., During, M.J., and Young, D. (2008). AAV vector-mediated RNAi of mutant huntingtin expression is neuroprotective in a novel genetic rat model of Huntington's disease. Mol. Ther. *16*, 947–956.
83. Leavitt, B.R., Raamsdonk, J.M., Shehadeh, J., Fernandes, H., Murphy, Z., Graham, R.K., Wellington, C.L., Raymond, L.A., and Hayden, M.R. (2006). Wild-type huntingtin protects neurons from excitotoxicity. J. Neurochem. *96*, 1121–1129.
84. Cattaneo, E., Rigamonti, D., Goffredo, D., Zuccato, C., Squitieri, F., and Sipione, S. (2001). Loss of normal huntingtin function: new developments in Huntington's disease research. Trends Neurosci. *24*, 182–188.
85. Nasir, J., Floresco, S.B., O'Kusky, J.R., Diewert, V.M., Richman, J.M., Zeisler, J., Borowski, A., Marth, J.D., Phillips, A.G., and Hayden, M.R. (1995). Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. Cell *81*, 811–823.
86. Zeitlin, S., Liu, J.P., Chapman, D.L., Papaioannou, V.E., and Efstratiadis, A. (1995). Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. Nat. Genet. *11*, 155–163.
87. Duyao, M.P., Auerbach, A.B., Ryan, A., Persichetti, F., Barnes, G.T., McNeil, S.M., Ge, P., Vonsattel, J.P., Gusella, J.F., Joyner, A.L., et al. (1995). Inactivation of the mouse Huntington's disease gene homolog Hdh. Science *269*, 407–410.
88. Dragatsis, I., Levine, M.S., and Zeitlin, S. (2000). Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. Nat. Genet. *26*, 300–306.
89. van Bilsen, P.H., Jaspers, L., Lombardi, M.S., Odekerken, J.C., Burright, E.N., and Kaemmerer, W.F. (2008). Identification and allele-specific silencing of the mutant huntingtin allele in Huntington's disease patient-derived fibroblasts. Hum. Gene Ther. *19*, 710–719.
90. Zhang, Y., Engelman, J., and Friedlander, R.M. (2009). Allele-specific silencing of mutant Huntington's disease gene. J. Neurochem. *108*, 82–90.